# Statistical Analysis of Solar Irradiance Variability

Angelos R. Nikolopoulos, Efstratios I. Batzelis, Paul Lewin
*School of Electronics and Computer Science*
*University of Southampton*
Southampton, UK
a.r.nikolopoulos@soton.ac.uk,e.batzelis@soton.ac.uk, pll@ecs.soton.ac.uk

Nikolaos Nikolaou
*Department of Physics and Astronomy*
*University College London*
London, UK
n.nikolaou@ucl.ac.uk

***Abstract*—Solar photovoltaic (PV) generation forecasting is an important tool to power system operators, but struggles under conditions of intermittent solar irradiance. Although studying and forecasting irradiance itself has been the subject of much research, little progress has been made on the variability (or fluctuation) of irradiance and its statistical properties, despite it being an important parameter in generation forecasting, state estimation and other power system applications. This paper takes a close look into the statistical nature of irradiance variability and shows that it can be sufficiently modeled by a Gaussian Mixture Model (GMM) of six components. Furthermore, an investigation on the required time resolution demonstrates that sub-minute resolution is necessary to accurately capture irradiance variability. The analysis is performed on a one-second resolution irradiance dataset provided by NREL.**

***Index Terms*—Gaussian Mixture Model (GMM), RR, solar irradiance, photovoltaic (PV) forecasting**

## I. Introduction

Solar photovoltaic (PV) power forecasting is hindered by volatile weather, especially in the presence of scattered and fast-moving clouds [1, 2]. Such conditions lead to highly intermittent and uncertain solar irradiance [3, 4], which translates to steep PV generation transients that can exceed half the nominal power within seconds [5, 6]. This high-frequency variation has been reported to negatively impact the electric grid in many ways, from power quality, such as voltage fluctuations and flicker, to stability, such as voltage dips and frequency oscillations [7, 8]. The field of solar forecasting has focused in the past on hourly-level resolution for energy yield; with the increasing levels of PV integration, however, it has become evident that higher time resolution is also important for operational reliability [9]. In this context, it is important to know not only the absolute value of solar generation, but also how quickly it changes, i.e. its variability (or fluctuation or ramping or ramp rate).

The research community has adopted the ramp rate (RR) metric to quantify the irradiance variability [8, 10, 11]. An analysis using RR to identify the short-term solar intermittency and its impact on a PV converter system [10] establishes the importance of solar variation in different timescales. Furthermore, the study [11] also employs RR to statistically characterize the irradiance variability that occurs during different weather patterns, demonstrating significant differences over multiple timescales. Study [8] uses the absolute value of RR to identify the temporal variability from one second to two minutes to investigate the short-term fluctuation at high altitudes, neglecting the negative RR values. Although these studies highlight the importance of time resolution on irradiance fluctuation, they do not quantify this impact and do not look into the statistics of irradiance variability.

National Renewable Energy Laboratories (NREL) was among the first to study the variability of Global Horizontal Irradiance (GHI) over time scales from 1 minute to 1 hour [11]. The study showed that the distribution of RR is very peaked (leptokurtic i.e. tall and skinny) due to the normal movement of the sun under constant sky conditions. This highly non-normal pattern was modeled in [11] via the Hyperbolic Distribution with moderate results. In contrast, this paper adopts Gaussian Mixture Models (GMM) for this task. GMM has been successfully used in the past in load profiling [12], in 15-minute PV output modeling [13] and GHI probability distribution [14], but this is the first time to use GMM in irradiance variability modeling.

In terms of time resolution, many studies have shown the importance of lower-than-hourly time intervals using minute-level data [11], but sub-minute resolutions have not been tested due to data unavailability. NREL raises awareness on the importance of second-level resolutions [1], but it remains unclear what the "optimal" resolution for accurate irradiance variability representation is. This is an important knowledge gap in light of the ultra-short forecasting/nowcasting of PV generation increasingly required in rich-solar networks.

This paper provides a holistic look into the irradiance variability by:

- Representing the irradiance RR at different weather using GMM for the first time
- Demonstrating that six GMM components offer the right balance between 'goodness-of-fit' and complexity
- Quantifying the impact of time resolution on RR and showing that few-seconds resolution is important

## II. Dataset

This study adopts the irradiance dataset from NREL [10], which comprises measurements from seventeen stations in Oahu that collected GHI data at 1-second intervals over the course of more than a year. The data from one of these stations, the AP3, was selected since it contains the highest level of robustness with the fewest erroneous values. It is used as a

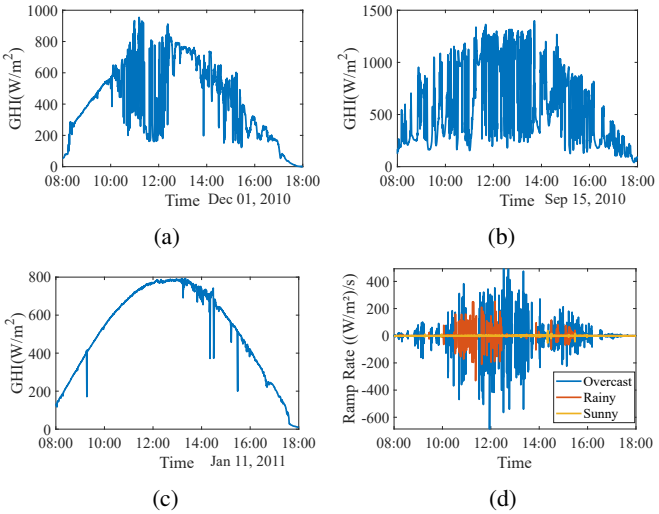



Fig. 1: GHI of the (a) rainy, (b) overcast and (c) sunny day. (d) Ramp Rate of these days.

benchmark, thereafter down-sampled at different resolutions to explore the impact of time step in RR modeling.

The weather pattern is a characteristic influencing factor for irradiance fluctuations, and therefore it needs to be considered in any such study. This paper adopts the weather classification proposed in [7], defining the weather as: *sunny* (minimal cloudiness), *overcast* (scattered clouds with high transparency gradients) and *rainy* (stochastic cloudiness with rain). Fig. 1(a)-(c) illustrate the irradiance time series for an indicative rainy, overcast and sunny day from the dataset used. The irradiance variability is strikingly different in the three weather types, quantified via the ramp rate in Fig. 1(d) showing up to 400 $Wm^{-2}s^{-1}$ in the overcast day.

## III. Statistical Distribution of Irradiance Variability

Throughout the literature, different metrics have been employed to quantify the irradiance/power variability, such as the rate of change, ramp rate, and rate of the ramp. This paper adopts the metric used in [10] defining the RR as the rate of change of irradiance within two consequent samples:

$$RR = \frac{d}{dt} GHI \tag{1}$$

Fig. 2 illustrates the RR of a probability density function (PDF) histogram of the rainy day in logarithmic y-axis (linear scale in the zoom box). As already observed in [11], the normal distribution cannot sufficiently represent such a tall and skinny trend caused by the sun movement during the constant-conditions times. The hyperbolic distribution used in [11] can deliver only minor improvement, as shown later.

This paper adopts the GMM theory for this representation, according to which the RR distribution $X$ can be modeled as a weighted sum of $k$ normal distributions $\phi(x_j,\ \theta_i)$, each

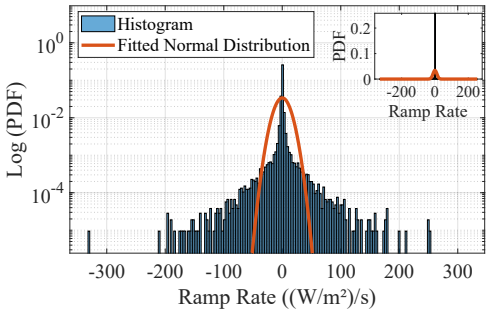



Fig. 2: Statistical distribution of RR on the rainy day.

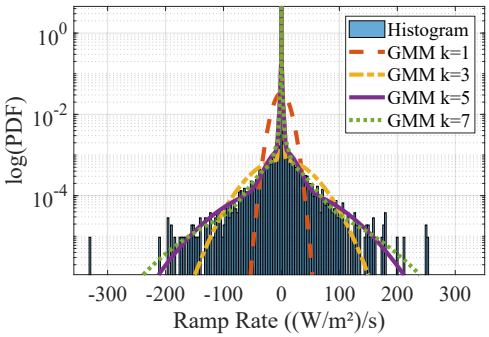



Fig. 3: GMM representation of RR on the rainy day with different components.

having different mean $\mu_i$ and variance $\sigma_i^2$.

$$f_X(x_j|\theta) = \sum_{i=1}^{k} w_i \phi(x_j,\ \theta_i), x_j \geq 0, j = 1, \ldots, N \tag{2}$$

$$\phi(x_j,\ \theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right) \tag{3}$$

$w_i$ is the weight of the $i$ GMM component, $N$ is the number of datapoints, and $\theta = (\{w_i,\ \mu_i,\ \sigma_i^2\}_{i=1}^{k})$.

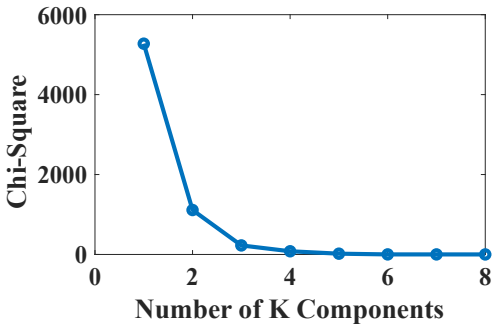
### A. Number of GMM Components

The number of GMM components $k$ required for this task is not a-priori known. Fig. 3 indicatively shows the GMM representation with 1, 3, 5 and 7 components for the rainy day; although 1 component (i.e. the standard normal distribution) is insufficient, more components lead to quite acceptable representations.

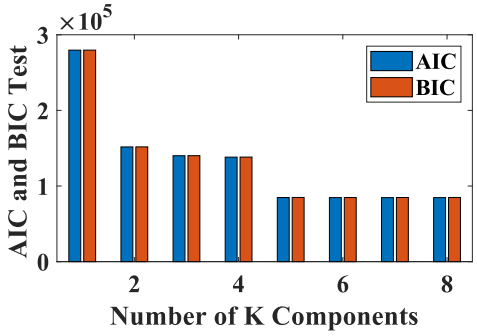
To measure the impact of the number of GMM components, the *chi-square goodness of fit* is adopted here, often used to evaluate statistical models [15]

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where $O_i$ refers to the $i$-th observed value and $E_i$ to the $i$-th expected value from the hypothesized distribution respectively. Fig. 4(a) plots the chi-square goodness-of-fit for up to 8 GMM components for the rainy day, showing clear modeling improvement with more components.




(a)




(b)

Fig. 4: Evaluation of GMM components for the rainy day. (a) Chi-Square Goodness of Fit and (b) AIC and BIC tests.

However, a high number of components may be unnecessary, and in fact undesirable in view of the added complexity entailed. To identify the number of components that strikes the right balance between accuracy and complexity, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used [16]:

$$AIC = -2\ln L + 2k \tag{5}$$

$$BIC = -2\ln L + \ln N \cdot k \tag{6}$$

$L$ refers to the likelihood but in this study we use (log)likelihood, $N$ refers to the number of data measurements and $k$ to the number of model parameters. Fig. 4(b) shows the AIC and BIC values for the rainy day, indicating that for $\geq 4$ components,the additional improvement in the quality of fit afforded may be outweighed by the added complexity.

Fig. 5 compares the 5-component GMM representation (GMM-5) to the Normal distribution and the Generalized Hyperbolic distribution based on [11]; GMM-5 is clearly more appropriate for RR modeling.

### *B. Analysis for all Weather Patterns and Yearly*

In this section, the steps performed for the rainy day are repeated for the overcast and sunny days, as well as for the *yearly* dataset that captures the RR over the course of an entire year. Although not a weather pattern, the yearly dataset reflects weather trends possibly not captured from the existing classification, and thus serves as a test dataset.

Fig. 6 shows the AIC test for different GMM components at all four case studies. Although the rainy, overcast and

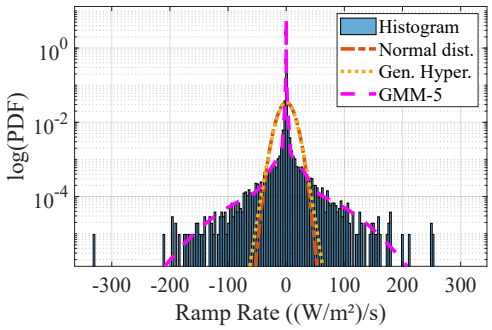



Fig. 5: Comparison of all models in RR representation for the rainy day.

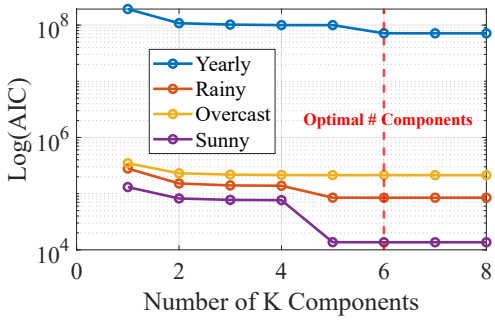



Fig. 6: AIC for different GMM components and weathers.

sunny days are sufficiently modeled via 5 components, the yearly dataset benefits from a 6th component. This is possibly due to the aforementioned unclassified trends not reflected in the three distinct weather patterns. In conclusion, selecting 6 GMM components is the safe choice for all weathers and patterns, as indicated by the red line in the plot.

The overall performance of the proposed GMM-6 model over the conventional Normal distribution and Generalized Hyperbolic for all case studies is given in Fig. 7. Clearly, GMM-6 is the most accurate by several orders of magnitude in the three distinct weather types, and by one order of magnitude in the yearly dataset.

## IV. Time Resolution Impact on Irradiance Variability

This section explores how the time resolution affects the RR in the sub-minute range. For these experiments, the original 1-second dataset was down-sampled at lower resolutions by averaging the irradiance values as in [17]. Artificial upscaling was studied by linear interpolating downsampled cases of 15s, 1min across 1s; subtracting them from the original 1s indicates an unreliable solution for high-resolution RR measurement.

Fig. 8 is an illustrative example of the lower resolution impact, demonstrating how much RR smooths out and decreases with 15 seconds and 1 minute sampling rates. Such visual examples have been sporadicly reported in the literature, but without quantification of the entailed accuracy loss.

An alternative way to illustrate this impact is via distribution curves, as in Fig. 9. This plot shows the RR distribution at various time resolutions from 1 second to 1 minute, by means

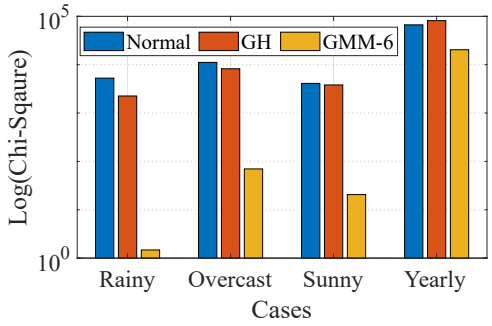



Fig. 7: Chi-square goodness of fit for all methods and case studies.

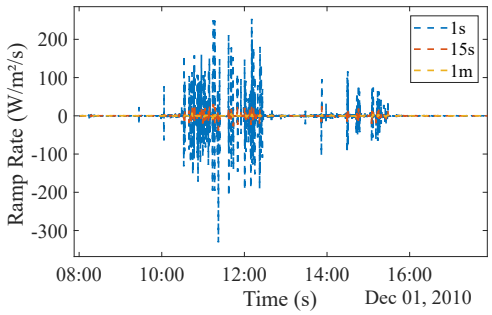



Fig. 8: RR time plot of the rainy day for different time resolutions.

of the respective fitted GMM-6 model. It is surprising that even slight down-sampling to 3 seconds reduces the distribution tail by 100 $Wm^{-2}s^{-1}$, i.e. almost by half. The misrepresentation is even higher at 15 seconds, but interestingly 45 seconds and 1 minute do not seem to differ much.

To further explore the time resolution impact on the shape of the RR distribution, the metrics of skewness and kurtosis are employed[18].

$$Skewness = \frac{\mu_3}{\sigma^3} \tag{7}$$

$$Kurtosis = \frac{\mu_4}{\sigma^4} \tag{8}$$

$\mu$ , refers to central moment and $\sigma$ to the standard deviation.

Fig. 10(a) shows the recorded skewness for the three weather patterns at various sub-minute resolutions. The rainy and overcast days exhibit near-zero skewness which indicates symmetrical RR distribution; the sunny day displays a more inconsistent behavior with skewness varying between positive and negative values. However, a closer look reveals that this is misleading due to the infrequent irradiance transients on the sunny day, i.e. small tails comprising few data points. Overall, the RR distribution is very symmetrical under all weather types and conditions, and this does not change with time resolution.

The kurtosis test at the same weather patterns and time resolutions is given in Fig. 10(b). The kurtosis values are always positive, which verifies that the RR distribution is leptokurtic, i.e. highly peaked, already known from [11]. However, Fig. 10(b) shows also how kurtosis reduces with lower time resolution, i.e. how less peaked the distribution becomes, which is another accuracy loss indication. Again,

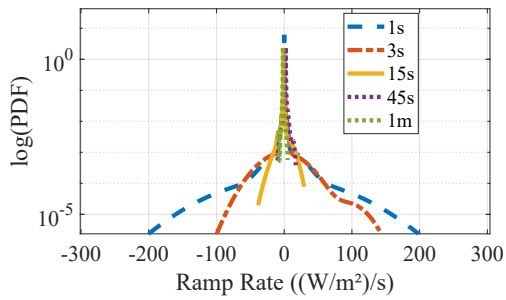



Fig. 9: RR distribution of the rainy day at different time resolutions.

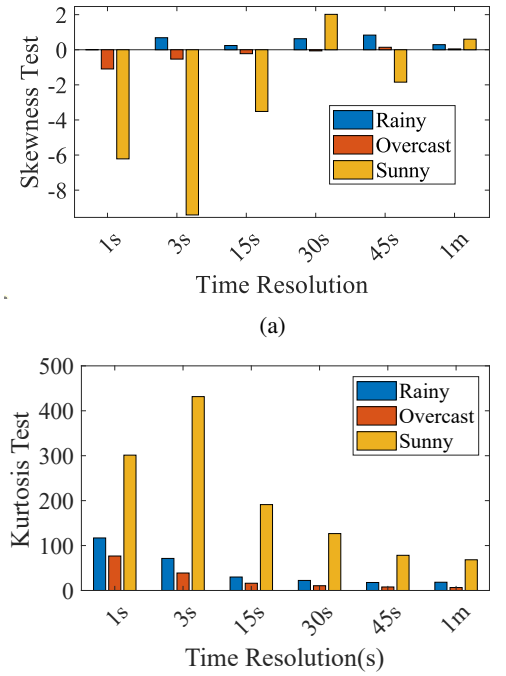



Fig. 10: (a) Skewness and (b) Kurtosis for the three weather patterns.

this modeling inaccuracy is evident even at 3-second down-sampling. The kurtosis increase from 1 second to 3 seconds in the sunny day is misleading for the abovementioned reasons.

The final test to quantify the impact of time resolution on the RR representation is via the chi-square goodness of fit. The GMM-6 model was fitted on down-sampled datasets (resolution of 5 seconds to 1 minute with increments of 5 seconds) and then evaluated against the original 1-second dataset via the chi-square goodness of fit. The results in Fig. 11 reach to an interesting conclusion: in all three cases,an accuracy loss of 2 orders of magnitude takes place in the first 5 seconds. After that, the sunny day loses less than 1 order of magnitude between 5 seconds and 1 minute, whereas the overcast and rainy days suffer about 6 and 10 orders of magnitude inaccuracy respectively. It is worth noting that these very high deviations arise from the way the chi-square goodness of fit is applied (GMM-6 models evaluated at different resolutions

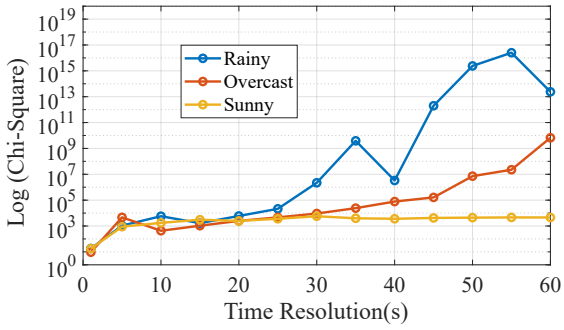



Fig. 11: Chi-square goodness of fit for different time resolutions.

than fitted on), and they do not entail the same level of impact on the grid: this remains to be explored. Nevertheless, the main conclusion is that the irradiance RR requires time resolution as low as 1 second; if this is not possible, then 1-minute resolution in sunny days is a reasonable compromise, but for any other weather pattern every second matters and sub-minute sampling is highly recommended.

## V. Conclusions

This paper shows that the distribution of irradiance variability can be adequately modeled via a GMM distribution and that six components is a valid selection for all weather types. The so-called GMM-6 model clearly outperforms alternatives in the literature. An investigation on time resolution also indicates that sampling periods of less than 5 seconds are required to entirely avoid information loss. Sunny days bear acceptable information loss at lower resolutions of 1 minute or more, but for overcast and rainy days sub-minute resolution is essential. Future work will look into ultra-short forecasting of irradiance variability and quantifying this impact on the electric grid.

## Acknowledgment

This work was supported by the Royal Academy of Engineering under the Engineering for Development Research Fellowship scheme (no. RF\201819\18\86).

## References

[1] M. Sengupta and A. Andreas, "Oahu solar measurement grid (1-year archive): 1-second solar irradiance; oahu, hawaii (data)."

[2] H. Jain, M. Sengupta, A. Habte, and J. Tan, "Quantifying solar pv variability at multiple timescales for power systems studies," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 2020, pp. 0180–0185.

[3] R. Blaga and M. Paulescu, "Quantifiers for the solar irradiance variability: A new perspective," *Solar Energy*, vol. 174, pp. 606–616, 2018.

[4] M. Paulescu, O. Mares, E. Paulescu, N. Stefu, A. Pacurar, D. Calinoiu, P. Gravila, N. Pop, and R. Boata, "Nowcasting solar irradiance using the sunshine number," *Energy Conversion and Management*, vol. 79, pp. 690–697, 2014.

[5] L. Liu, Y. Zhao, D. Chang, J. Xie, Z. Ma, Q. Sun, H. Yin, and R. Wennersten, "Prediction of short-term pv power output and uncertainty analysis," *Applied Energy*, vol. 228, pp. 700–711, 2018.

[6] M. Paulescu, E. Paulescu, and V. Badescu, "Chapter 9 - nowcasting solar irradiance for effective solar power plants operation and smart grid management," in *Predictive Modelling for Energy Management and Power Systems Engineering*, R. Deo, P. Samui, and S. S. Roy, Eds. Elsevier, 2021, pp. 249–270.

[7] D. Niu, K. Wang, L. Sun, J. Wu, and X. Xu, "Short-term photovoltaic power generation forecasting based on random forest feature selection and ceemd: A case study," *Applied Soft Computing*, vol. 93, p. 106389, 2020.

[8] I. Ranaweera, O.-M. Midtgård, and G. H. Yordanov, "Short-term intermittency of solar irradiance in southern norway," in *29th European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC)*, 2014, pp. 2635–2638.

[9] M. Lave, J. Kleissl, and E. Arias-Castro, "High-frequency irradiance fluctuations and geographic smoothing," *Solar Energy*, vol. 86, no. 8, pp. 2190–2199, 2012, progress in Solar Energy 3.

[10] Y. Yao, N. Ertugrul, and A. P. Kani, "Investigation of short-term intermittency in solar irradiance and its impacts on pv converter systems," in *2022 32nd Australasian Universities Power Engineering Conference (AUPEC)*, 2022, pp. 1–6.

[11] B. M. Hodge, M. Hummon, and K. Orwig, "Solar ramping distributions over multiple timescales and weather patterns."

[12] R. Singh, B. C. Pal, and R. A. Jabr, "Statistical representation of distribution system loads using gaussian mixture model," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 29–37, 2010.

[13] Z. Wang, J. Kang, L. Cheng, Z. Pei, C. Dong, and Z. Liang, "Mixed gaussian models for modeling fluctuation process characteristics of photovoltaic outputs," *Frontiers in Energy Research*, vol. 7, 2019.

[14] M. Wahbah, T. H. M. EL-Fouly, and B. Zahawi, "Gaussian mixture model for estimating solar irradiance probability density," in *2020 IEEE Electric Power and Energy Conference (EPEC)*, 2020, pp. 1–6.

[15] W. G. Cochran, "The 2 test of goodness of fit," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 315–345, 1952.

[16] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[17] A. Gagné, D. Turcotte, N. Goswamy, and Y. Poissant, "High resolution characterisation of solar variability for two sites in eastern canada," *Solar Energy*, vol. 137, pp. 46–54, 2016.

[18] L. Garcia-Gutierrez, C. Voyant, G. Notton, and J. Almorox, "Evaluation and comparison of spatial clustering for solar irradiance time series," *Applied Sciences*, vol. 12, no. 17, 2022.